Dimensionality reduction in exposome and omic data to uncover molecular pathologic mechanisms

Carlos Ruiz-Arenas

Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona, Spain; Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain

Carles Hernandez-Ferrer

Centro Nacional de Análisis Genómico (CNAG-CRG); Center for Genomic Regulation (CRG); Barcelona Institute of Science and Technology (BIST), Barcelona, Catalonia, Spain

♦ The following is a project proposal for the Exposome Data Challenge 2021 ♦

Keywords: exposome, transcriptome, metabolome, dimensionality reduction, pathway association analyses, integration analyses.

Background: The exposome represents the sum of all environmental exposures over a lifetime and, at the intersection of public health and toxicology, is a key tool to look forward in the understanding of the disease development machinery, completing the knowledge obtained from genetics (1). The traditional genome-wide association studies inspired a range of methods to investigate the underlying links in exposome with multiple health outcomes (2). The exposome-wide association approach consists of a series of linear and logistic regression models to assess the association between single exposures and outcomes (3). Current exposome studies have included omic measurements (aka, trascriptomics, proteomics, and metabolomics) in the study design. Omic measurements have provided new insights for disease physiology and have been proved as valuable biomarkers for disease (4,5). However, omic measurements greatly increase the data dimensionality, requiring enormous datasets to achieve statistical power to detect associations using single exposure-omic measurements tests. In addition, omic features have not typically a clear biological meaning (6), so exposure-omic features associations reaching statistical significance are commonly difficult to interpret and translate to biological knowledge.

Goal: We aim to define a new approach to elucidate the relationships between the exposome and a series of molecular signatures (aka, trascriptomics, proteomics, and metabolomics) in a generalized fashion. To improve the interpretability of the results, we will reduce the data dimensionality by mapping the exposures and molecular signatures to biochemical and biological pathways.

Methods: We will collapse the exposume into biochemical pathways using an exposure-to-exposure mapping from the *Human Metabolome Database* (10) and the *Toxin and Toxin-Target Database* (11). Same methodology will be applied to both urine and serum metabolome datasets. We will consider merging urine and serum datasets, after collapsing them to pathway level. We will collapse gene expression in pathways activation using the *hipatia* (4) algorithm. Then, the state of the art exposome-wide association analysis will be performed but at pathway-to-pathway level. Finally, *MOFA+* (7,9) and other integrative methodology (5,8,12–15) may be considered to corroborate the previous results.

Expected results: First set of results comprehends the relation between each feature (exposome and omic data) with the pathway(s) it belongs to (results of dimensionality reduction) and the cross-association between the exposome and the molecular signatures (results from the exposome-wide association study at pathway level). If considered, the second set of results, from the integration analysis, provide corroboration of the relationship exposure-to-molecular signature in an N-to-N consideration.

Challenges: 1) *Combined effects of exposures*, thanks to dimensionality reduction on the exposome side; 2) *Using omics data to improve inference on the link between exposome and health*, due to the exposome-wide association study at pathway level between the reduced exposome and reduced gene expression and metabolome; and 3) *Multi-omics analysis*, due to the methods for data integration that will include the exposome, the gene expression, and the metabolome.

Bibliography:

1. Wild CP. The exposome: from concept to utility. Int J Epidemiol. 2012 Feb;41(1):24–32.
2. Fallin MD, Kao WHL. Is "X"-WAS the future for all of epidemiology? Epidemiology. 2011 Jul;22(4):457–9; discussion 467-468.
3. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS One. 2010 May 20;5(5):e10746.
4. Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J. High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget. 2017 Jan 17;8(3):5160–78.
5. Fan Z, Zhou Y, Ressom HW. MOTA: Network-Based Multi-Omic Data Integration for Biomarker Discovery. Metabolites. 2020 Apr 8;10(4).
6. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nat Rev Genet. 2017 Jul;18(7):441–51.
7. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology. 2018 Jun 1;14(6):e8124.
8. Reinke SN, Galindo-Prieto B, Skotare T, Broadhurst DI, Singhania A, Horowitz D, et al. OnPLS-Based Multi-Block Data Integration: A Multivariate Approach to Interrogating Biological Interactions in Asthma. Anal Chem. 2018 Nov 20;90(22):13400–8.
9. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology. 2020 May 11;21(1):111.
10. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Research. 2018 Jan 4;46(D1):D608–17.
11. Lim E, Pon A, Djoumbou Y, Knox C, Shrivastava S, Guo AC, et al. T3DB: a comprehensively annotated database of common toxins and their targets. Nucleic Acids Res. 2010 Jan;38(Database issue):D781-786.
12. Csala A, Hof MH, Zwinderman AH. Multiset sparse redundancy analysis for high-dimensional omics data. Biom J. 2019 Mar;61(2):406–23.
13. Park M, Kim D, Moon K, Park T. Integrative Analysis of Multi-Omics Data Based on Blockwise Sparse Principal Components. Int J Mol Sci. 2020 Nov 2;21(21).
14. Srivastava V, Obudulu O, Bygdell J, Löfstedt T, Rydén P, Nilsson R, et al. OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI- superoxide dismutase Populus plants. BMC Genomics. 2013 Dec 17;14(1):893.
15. Csala A, Zwinderman AH, Hof MH. Multiset sparse partial least squares path modeling for high dimensional omics data analysis. BMC Bioinformatics. 2020 Jan 9;21(1):9.